VU Research Portal

Maximum likelihood estimation by monte carlo simulation

Peng, Yijie; Fu, Michael C.; Heidergott, Bernd; Lam, Henry

Link to publication in VU Research Portal

citation for published version (APA)
Peng, Y., Fu, M. C., Heidergott, B., & Lam, H. (2020). Maximum likelihood estimation by monte carlo simulation:
Toward data-driven stochastic modeling. *Operations Research*, *68*(6), 1896-1912.
https://doi.org/10.1287/opre.2019.1978

Operations Research

Maximum Likelihood Estimation by Monte Carlo Simulation: Toward Data-Driven Stochastic Modeling

Yijie Peng, Michael C. Fu, Bernd Heidergott, Henry Lam

Methods

Maximum Likelihood Estimation by Monte Carlo Simulation: Toward Data-Driven Stochastic Modeling

Yijie Peng,[a] Michael C. Fu,[b] Bernd Heidergott,[c] Henry Lam[d]

[a] Department of Management Science and Information Systems, Guanghua School of Management, Peking University, 100871 Beijing, China;
[b] Robert H. Smith School of Business and Institute for Systems Research, University of Maryland, College Park, Maryland 20742;
[c] Department of Econometrics and Operations Research, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands; [d] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
Contact: pengyijie@pku.edu.cn, https://orcid.org/0000-0003-2584-8131 (YP); mfu@umd.edu (MCF); b.f.heidergott@vu.nl (BH); henry.lam@columbia.edu (HL)

Abstract. We propose a gradient-based simulated maximum likelihood estimation to estimate unknown parameters in a stochastic model without assuming that the likelihood function of the observations is available in closed form. A key element is to develop Monte Carlo–based estimators for the density and its derivatives for the output process, using only knowledge about the dynamics of the model. We present the theory of these estimators and demonstrate how our approach can handle various types of model structures. We also support our findings and illustrate the merits of our approach with numerical results.

1. Introduction

Maximum likelihood estimation (MLE) is the most popular statistical technique for estimating unknown parameters based on sample observations. In addition to its dominant role in parameter estimation, MLE also plays important roles in other inference problems, such as hypothesis testing and model selection. Under mild regularity conditions, MLE has the following desirable properties: consistency, asymptotic normality, asymptotic efficiency (achieving the Crámer–Rao bound), and functional invariance (Shao 2003).

A basic requirement for applying MLE is the availability of the likelihood function. In common statistical models, the likelihood function of the data can be written analytically, and thus, MLE can be naturally applied. In this paper, we focus on stochastic models represented by system dynamics rather than likelihood functions, motivated by scenarios of the following type. An idealized stochastic model—think, for example, of a queueing model—is given, and a fixed set of data output from the "real" system (e.g., the system times of the customers) is available. Then, MLE is used to select model parameters (e.g., arrival and service rates) that maximize the likelihood function under the idealized model from the given data. Intuitively, this

maximizes the "agreement" of the selected model parameters with the observed data. However, in the typical case in which the stochastic model is represented via dynamic equations (e.g., Lindley's recursion), the likelihood function of the output data may be unavailable in analytical form, a challenge that we investigate in this paper. In other words, our work focuses on the *computational* aspects of MLE using a simulation-based method rather than the statistical properties of MLE.

Our motivation is twofold. First, in building stochastic models for various analyses, the modeler needs to calibrate parameters, for example, interarrival and service rates in a queue. When system input data (e.g., interarrival and service times) are readily available, MLE (or other estimation techniques) can easily be performed on the input data. However, there are practical scenarios in which, because of data collection or operational constraints, only output-level data are available. Inferring input parameters from output data are generally known as an "inverse problem" and has been studied in several scientific areas (Kennedy and O'Hagan 2001, Tarantola 2005). Similar ideas have also been used in studying economic market structure, often called inverse optimization (Bertsimas et al. 2012, Birge et al. 2017, Esfahani et al. 2018). In stochastic

modeling, the inverse problem has been investigated in Basawa et al. (1996, 2008), Pickands and Stine (1997), Fearnhead (2004), Wang et al. (2006), and Ross et al. (2007) and is also related to the so-called "queue inference engine" in Larson (1990) and point process approximations in Whitt (1982). This literature exploits closed-form representations of (typically) queueing models or approximations, such as heavy-traffic limits. Our approach is a general technique that relies on knowledge of the underlying dynamics of the stochastic model, using simulation and gradient estimation to perform MLE rather than closed-form approximations. A related work is Goeva et al. (2019) that also considers simulation-based calibration but from a different perspective of robust optimization.

Second, we take the viewpoint that, when the input model or the dynamic is potentially misspecified and, as such, the statistical consistency in output prediction no longer holds, it could be beneficial to fit the output data instead of the input data. In general, MLE is an asymptotic minimizer of the Kullback–Leibler (KL) divergence between the conjectured model and the data (Van der Vaart 2000). Thus, applying MLE on the output level attempts to minimize the statistical discrepancy between models and data when the output prediction accuracy is important, which is often the case when building stochastic models. This idea of "best fitting" at the output level is similar to the training of machine learning algorithms, which in recent years have been developed to find reliable representations of observed (output) data by statistical (econometric) models. Although such approaches have merit in finding simplified representations of high-dimensional data, they lead to a black-box model rather than a *causal* representation provided by stochastic modeling. ("We bring light to the black box.")

The MLE method developed in this paper requires that a class of parameterized causal models, which can be analyzed through simulation, be given. The main technical contribution of this work is that we derive unbiased estimators for the density and its derivatives for the output of a generic stochastic model by Monte Carlo simulation. The likelihood of the (output) data is the joint density evaluated at the observations. For a continuous distribution, we write the density as the derivative of the distribution function, which can be viewed as the expectation of an indicator function. Thus, deriving an unbiased estimator for estimating the density requires addressing the discontinuity introduced by the indicator function and the structural parameters in the sample performance. Infinitesimal perturbation analysis (IPA) cannot deal with discontinuities, and the likelihood ratio (LR) method cannot handle structural parameters in the sample performance (Fu 2015).

The generalized likelihood ratio (GLR) method in Peng et al. (2018) can deal with a larger scope of discontinuities in the sample performance. We use this technique to estimate the density and its derivatives, which fall under the umbrella of "distribution sensitivities"—derivatives of the distribution function with respect to (w.r.t.) both the argument and parameter in an underlying stochastic model—in Lei et al. (2018). The difficulty in distribution sensitivities lies in the discontinuity in the sample performance and the presence of structural parameters. Previously, Hong and Liu (2010) offered a pathwise derivative estimator w.r.t. the parameter in the underlying stochastic model, which achieves a convergence rate slower than the canonical square-root rate in Monte Carlo simulation, whereas the GLR estimator achieves the square-root convergence rate even for higher-order derivatives (Glynn et al. 2020a).

With the GLR estimators for the density and its derivatives, we propose a general gradient-based simulated maximum likelihood estimation (GSMLE) method, which applies a stochastic approximation (SA) algorithm to estimate the unknown parameters in stochastic models without assuming an analytical form of the likelihood function. GSMLE can deal with independent and identically distributed (i.i.d.) observations and also data generated by a Markovian model, for example, system times of a G/G/1 queue, and more generally hidden Markov models (HMM). In the latter context, related work includes Peng et al. (2014, 2016), who calibrate stochastic volatility (SV) models using MLE that indicate computational advantages over some benchmark methods in Bayesian estimation and moment estimation. The discrete observations of the SV models can be viewed as the observable states of an HMM. Because of the presence of the hidden states, simulation is implemented to estimate the likelihood and its derivatives, but these previous works assume the observational kernel associated with the HMM and its derivatives are analytically known or can be numerically calculated by Fourier inversion, which is not assumed in this work.

We summarize our main contributions as follows:
• We propose a new method for estimating unknown parameters of a stochastic model without assuming an analytical likelihood function.
• We directly fit the underlying stochastic model to the output data, which opens the possibility of extending data-driven ideas to causal stochastic models.
• We generalize our scheme to efficiently utilize the simulated samples in calculating the MLE for an HMM.

The rest of this paper is organized as follows. In Section 2, we formulate the problem and provide illustrative examples. We propose the GSMLE in Section 3, in which the distribution sensitivity estimators

are derived in Section 3.1, and the sample path derivative estimator for the likelihood of the HMM is given in Section 3.2. Numerical results can be found in Section 4. Conclusions are given in the last section. Some of the proofs and additional numerical experiments can be found in the online appendix.

2. Problem Formulation

The MLE for parameter θ governing a given parametric family of stochastic models is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_T(\theta),$$

where Θ is the feasible set for parameter θ and $L_T(\theta)$ is the log-likelihood of observations. Throughout the paper, we assume that the distribution of the observations is continuous and admits a density. However, the likelihood and density are not available in analytical form, let alone the MLE.

We first study the case of i.i.d. observations from a data-generating process specified by a stochastic model. More specifically, for $t = 1, \ldots, T$, we let

$$Z_t = g(X_t; \theta),$$

where $X_t = (X_{1,t}, \ldots, X_{n,t})$, $t = 1, \ldots, T$, represent the i.i.d. input random variables (r.v.s) with a given joint density $f(x; \theta)$, and the mapping $g(\cdot, \cdot)$ maps the input to the observable output Z_t, thus representing the stochastic model. The parametric forms of $f(\cdot; \cdot)$ and $g(\cdot, \cdot)$ are assumed known and amenable to simulation, but the parameter θ is unknown and needs to be inferred from the output observations Z_t, which can be a vector. The log-likelihood of the output Z_t is given by

$$L_T(\theta) = \sum_{t=1}^{T} \log p(Z_t; \theta),$$

where $p(\cdot; \theta)$ is the density of Z_t that lacks an analytical form in general. The asymptotic properties of the MLE for i.i.d. observations are well known and can be found in Shao (2003).

If X_t is one-dimensional with density p_{X_t}, then, provided that g is invertible and the inverse is differentiable with respect to the z argument, it follows from a standard result in probability that the density of Z_t can be obtained in closed form as

$$p(z; \theta) = p_{X_t}\left(g^{-1}(z; \theta)\right)\left|\frac{d}{dz}g^{-1}(z; \theta)\right|,$$

and the log-likelihood is given by

$$L_T(\theta) = \sum_{t=1}^{T} \log\left(p_{X_t}\left(g^{-1}(Z_t; \theta)\right)\left|\frac{d}{dz}g^{-1}(z; \theta)\right|_{z=Z_t}\right).$$

The key condition that limits the applicability of this direct approach is the requirement of the invertibility of g. The theory developed in this paper does not require this restrictive property and only requires g to be differentiable with respect to x and its derivative to be nonzero almost everywhere (a.e.). Moreover, the approach developed in this paper works in the case of multidimensional input r.v. X_t as well. The following example illustrates the generality of our approach in the one-dimensional case.

Example 1. Let X_t have support $[0, \infty)$ with density $p_{X_t}(x)$ and consider $Z_t = \theta \sin(X_t)$. Note that $g(x; \theta) = \theta \sin(x)$ fails to be invertible on the support of X_t. However, g is differentiable with respect to x and θ and $\partial g(x; \theta)/\partial x \neq 0$ a.e. As we discuss later on, this makes it possible to apply our estimator.

An extension of the i.i.d. case is to assume observations come from a data-generating process following a Markov chain. The analysis of Markov chains is addressed in the subsequent section.

2.1. Markov Chains

We assume that Markov chain $\{Z_t : 0 \leq t \leq T\}$ is driven by the following stochastic recursion:

$$Z_t = g(X_t; Z_{t-1}, \theta), \tag{1}$$

where $X_t, t = 1, \ldots, T$, are i.i.d. (input) r.v.s driving the Markov chain and Z_0 is the initial state independent of θ. The log-likelihood of observations following a Markov process is given by

$$L_T(\theta) = \sum_{t=1}^{T} \log p(Z_t; Z_{t-1}, \theta),$$

where $p(\cdot; Z_{t-1}, \theta)$ is the (unknown) conditional density on Z_{t-1}.

Example 2. Customers arrive at a service station according to a renewal point process. The interarrival times $\{A_t : t \in \mathbb{N}\}$ are i.i.d. with density $f_A(x)$ and $0 < E[A_t] < \infty$ and $\mathbb{P}(A_t = 0) = 0$. Customers are served in order of arrival, and consecutive service times are i.i.d. random variables $\{B_t(\theta) : t \in \mathbb{N}\}$ with density $f_B(x; \theta)$. Interarrival times and service times are assumed to be mutually independent. Consider the process of consecutive sojourn (or system) times $\{Z_t(\theta)\}$, denoting the total time that the corresponding customer is in the system (from arrival to end of service). The arrival process starts at $T_0 = 0$. Consecutive sojourn times $Z_t(\theta)$ follow the well-known Lindley equation:

$$Z_t(\theta) = \max(0, Z_{t-1}(\theta) - A_t) + B_t(\theta), \quad t \geq 1, \tag{2}$$

where we assume that the system starts empty and formally set $Z_0(\theta) = 0$. Letting $X_t = (A_t, B_t(\theta))$, mapping (1) becomes

$$g(x_t; z) = \max(0, z - a_t) + b_t.$$

The mapping g is differentiable with respect to b_t everywhere and differentiable a.e. with respect to a_t. We show the MLE of this model can be estimated by our method.

Complex stochastic systems are typically difficult to accurately model, which leads to a model misspecification problem. Let $p(\cdot)$ be the density of output r.v. Z of the true model and $\{p_\theta(\cdot) : \theta \in \Theta\}$ be the corresponding density of the output r.v. of a misspecified parametric family, that is, $p(\cdot) \notin \{p_\theta(\cdot) : \theta \in \Theta\}$. Let $\tilde{\theta}$ be a solution that minimizes the KL divergence:

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} E\big[\log(p(Z)/p_\theta(Z))\big]. \tag{3}$$

It is easy to show that the MLE $\hat{\theta}$ is consistent with $\tilde{\theta}$ as the sample size goes to infinity, that is, $\hat{\theta} - \tilde{\theta} \to 0$ as $T \to \infty$ (Van der Vaart 2000). For a given parametric family of potential models, property (3) of the MLE allows for finding the model within the given family that best explains the observed output data. It is worth noting that the meaning of "best fit" is relative to the assumed parametric family. For example, we may want to fit an (oversimplified) queue, such as the M/M/1 queue, to the output data of a general G/G/1 queueing system. By fitting an M/M/1 to output data of a G/G/1 queue, we drive the statistical property of the outputs of two systems closer, leading to the best, although misspecified, explanatory M/M/1 model for the observed data. We illustrate the impact of model misspecification and the use of MLE in compensating for a possible model misfit with the following variation of Example 2.

Example 3 (Revisit Example 2). We apply the MLE to this example; however, the arrival process of the true model is a Markov modulated process (MMP). More specifically, let η_t be a Markov chain (P_{ij}) on $\{1, 2\}$. The sequence of interarrival times is given by $A_t = \zeta_{t,\eta_t}$, where $\zeta_{t,1}$ and $\zeta_{t,2}$ are independent random variables. For our experiment, we let

$$P_{11} = P_{22} = 1 - \eta, \quad P_{12} = P_{21} = \eta,$$

for $\eta \in (0, 1)$. The arrival process is, thus, an MMP. Applying the MLE assuming the model in Example 2 (the misspecified model) with the output data from the preceding model (the "reality") provides a mechanism to (potentially better) calibrate the misspecified model. Later on, we present supporting numerical examples.

2.2. Hidden Markov Models

A further extension to the Markov chain framework is that observations follow an HMM. HMMs have wide applications, including pattern recognition, genetic engineering, clustering analysis, and finance (Dymarski 2011). Because of the presence of hidden states, the likelihood of the HMM is a high-dimensional integral. Estimating the likelihood itself is a challenge, let alone maximization. Thus, statistical inference has been a central part of HMM research (Cappé et al. 2005). In statistics, the expectation-maximization (EM) algorithm is a popular method to calculate the MLE with latent variables (Dempster et al. 1977). The EM algorithm separates estimating the expectation (simulation) from maximizing the parametric performance, which requires iteratively solving a series of optimization problems, and the computational benefit of the EM algorithm relies heavily on assuming an exponential family for the joint likelihood of hidden and observable states. In contrast, GSMLE optimizes the likelihood simultaneously through simulation optimization and estimates the observational kernel and its derivatives by Monte Carlo simulation rather than assuming an analytical form.

An HMM can be specified by the following general state space model: for $t = 1, \ldots, T$,

$$Z_t = g(X_t; S_t, \theta), \quad S_t = h(Y_t; S_{t-1}, \theta), \tag{4}$$

where Y_t, $t = 1, \ldots, T$, are i.i.d. r.v.s driving the (hidden) underlying Markov chain $\{S_t\}$ with initial state S_0 independent of θ, and X_t, $t = 1, \ldots, T$, are i.i.d. r.v.s introducing interference to the (unobservable) state S_t of the Markov chain. Only Z_t, $t = 1, \ldots, T$, are observable. Put differently, we only observe a noisy signal from the underlying system. For given observation data $Z_1, \ldots, Z_T$, the log-likelihood of observations following an HMM is given by

$$L_T(\theta) \doteq \log E\left[\prod_{t=1}^{T} p(Z_t; S_t, \theta)\right], \tag{5}$$

where $p(\cdot; S_t, \theta)$ is the (unknown) conditional density of observation Z_t on hidden state S_t, which is also called the *observational kernel*, and the expectation operator is applied to average out S_t. The asymptotic properties of the MLE for an HMM are similar to the i.i.d. case and can be found in Cappé et al. (2005, chapter 6).

Example 4. We now discuss an application in which θ serves as a behavioral threshold parameter. We adjust the notation introduced in Example 2 by letting S_t denote the sojourn time of the tth customer. Once a customer finishes service, the customer is asked about the rating of the service, which is a summation of some random factor X_t and the quality of the service measured by $c_1(\theta - S_t)$ if $S_t \le \theta$ and $c_2(\theta - S_t)$ if $S_t > \theta$ with $0 < c_1 < c_2 < \infty$. Note that X_t models the part of the rating that cannot be explained by our simplified model. Assuming $c_1 < c_2$ reflects the asymmetric perception of gain and loss in human behavior. This models the

rating of the service depending on the behavioral parameter θ as follows:

$$g(X_t; S_t, \theta) = c_1 \max(\theta - S_t, 0) + c_2 \min(\theta - S_t, 0) + X_t,$$

where

$$S_t = \max(0, S_{t-1} - A_t) + B_t.$$

Provided we observe the rating of the customers, we can apply MLE to identify the choice for the behavioral parameter in our utility model.

3. GSMLE Theoretical Development

In our paper, we solve the MLE by gradient-based simulation optimization. Specifically, the following SA algorithm is used:

$$\theta_{k+1} = \Pi_\Theta\left[\theta_k + \lambda_k \widehat{D}_T(\theta_k)\right], \tag{6}$$

where Π_Θ is a projection onto Θ with Θ denoting a given compact set of admissible parameters, $\widehat{D}_T(\theta)$ is the log-likelihood derivative estimator at θ, and λ_k is the step size at iteration k. To guarantee almost sure (a.s.) convergence of SA to the optimum of the log-likelihood, certain conditions on the noise of the derivative estimate, the sequence of step sizes, and uniqueness of the optimum are required; see Kushner and Yin (2003, chapter 5) for details.

Unlike random search, gradient-based simulation does not suffer from the curse of dimensionality (of parameter θ), and it is usually considered to be efficient if it applies. For objective functions with multiple local optima, SA can be implemented with different initializations, and the best terminal estimate is chosen. In this paper, we use the SA configurations suggested by the conventional Robbins–Monro algorithm (Kushner and Yin 2003). In the next section, we provide unbiased derivative estimators for the density of the observations. The resulting log-likelihood derivative estimator has a ratio form that bears a small bias relative to the standard deviation, which can be reduced by increasing the number of simulation replications.

3.1. Distribution Sensitivities

For simplicity, we only consider distributional sensitivities when θ is scalar (X_t is still a vector). It is straightforward to extend the results in this section to the case in which θ is a vector. We first derive the distribution sensitivity estimators for the i.i.d. case. The likelihood of the observations and its sensitivities are given as follows:

$$p(Z_t; \theta) = \frac{\partial \, \mathrm{E}\left[\mathbf{1}\{g(X_t; \theta) \leq z\}\right]}{\partial z}\Bigg|_{z=Z_t},$$

$$\frac{\partial p(Z_t; \theta)}{\partial \theta} = \frac{\partial^2 \, \mathrm{E}\left[\mathbf{1}\{g(X_t; \theta) \leq z\}\right]}{\partial \theta \partial z}\Bigg|_{z=Z_t}. \tag{7}$$

Notice that expectation $\mathrm{E}[\mathbf{1}\{g(X_t; \theta) \leq z\}]$ is the distribution function of Z_t. To estimate the distributional sensitivities in (7), note that (i) IPA might not apply because of the discontinuity introduced by the indicator function in the sample performance, and (ii) LR might not work because of the presence of structural parameters, that is, θ and z, in the sample performance. In the following, we show that our GLR estimator overcomes both (i) and (ii).

Let X follow the distribution of X_t, $t = 1, \ldots, T$ with density $f(x; \theta)$, where $x \doteq (x_1, \ldots, x_n)$. We derive n different GLR estimators, one utilizing each component of the input vector (but each only using one simulation run) for the derivative of the distribution function with respect to its argument (i.e., the density $p(z; \theta)$ in (7)) in the form

$$\mathbf{1}\{g(X; \theta) \leq z\}\, \Psi_{1,i}(X; \theta), \quad i = 1, \ldots, n, \tag{8}$$

where $\Psi_{1,i}(x; \theta)$ can be expressed explicitly in terms of the derivatives of $f(\cdot; \theta)$ and $g(\cdot; \theta)$ (shown before Theorem 1). We also derive an estimator for the second-order distribution sensitivity w.r.t. θ and z (i.e., $\partial p(z; \theta)/\partial \theta$ in (7)) as

$$\mathbf{1}\{g(X; \theta) \leq z\}\, \Psi_{2,i}(X; \theta), \quad i = 1, \ldots, n,$$

where $\Psi_{2,i}(X; \theta)$ is expressible in terms of both $\Psi_{1,i}(x; \theta)$ and the derivatives of $f(\cdot; \theta)$ and $g(\cdot; \theta)$ (shown before Theorem 2).

With these estimators of the density and its derivative, we can estimate the derivative of the log-likelihood in the i.i.d. setting:

$$\frac{\partial L_T(\theta)}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial p(Z_t; \theta)}{\partial \theta}\left(p(Z_t; \theta)\right)^{-1},$$

by plugging the corresponding estimators into the numerator and denominator of the ratio

$$\widehat{D}_{T,i}(\theta) = \sum_{t=1}^{T}\left(\sum_{m=1}^{M}\mathbf{1}\left\{g\left(X_t^{(m)}; \theta\right) \leq Z_t\right\}\Psi_{2,i}\left(X_t^{(m)}; \theta\right)\right)$$
$$\times \left(\sum_{m=1}^{M}\mathbf{1}\left\{g\left(X_t^{(m)}; \theta\right) \leq Z_t\right\}\Psi_{1,i}\left(X_t^{(m)}; \theta\right)\right)^{-1},$$

where M is the number of simulated samples and $X_t^{(m)}$ is the mth sample of X_t. The Markovian case can be handled similarly by replacing the density with the conditional density (transition kernel) of the Markov chain. Therefore, we omit the details on the Markovian case for brevity. Then, SA (6) can be used to search for the optimum of the likelihood function.

The first-order distribution sensitivity with respect to z discussed previously (and also the sensitivity with respect to θ) is pivotal in quantile sensitivity estimation pioneered by Hong (2009), a seminal work leading to a series of work on sensitivity estimation

for financial risk measures (Fu et al. 2009, Hong and Liu 2009, Liu and Hong 2009, Jiang and Fu 2015, Heidergott and Volk-Makarewicz 2016). Recently, Peng et al. (2017) and Glynn et al. (2020b) established the asymptotic results for several quantile sensitivity estimators in a unified manner using functional limit theory.

For our main theoretical results, Theorem 1 presents the GLR estimator for the first-order distribution sensitivity with respect to z, which is a special case of the general setting in Peng et al. (2018) under simpler conditions (A.1)–(A.3) enabled by an explicit construction of a smoothing sequence for the performance function that utilizes the specific structure of the indicator function. Theorem 2 extends the GLR estimator, for the first time, to any order of the distribution sensitivity. This, in particular, is used to estimate $\partial p(z; \theta)/\partial \theta$.

Define $A_{z,\theta}^{\epsilon} \doteq \{x \in \mathbb{R}^n : z - \epsilon \leq g(x; \theta) \leq z + \epsilon\}$. We introduce the following regularity conditions to derive the GLR estimator for the first-order distribution sensitivity in (8) with index i.

Condition A.1. *Suppose the components of the random vector X_t are independent, that is, $f(x; \theta) = \prod_{l=1}^{n} f_l(x_l; \theta)$, $f(x; \theta)$ is differentiable and $g(x; \theta)$ is twice differentiable on $\mathbb{R}^n \times \Theta$.*

Condition A.2. *The following uniform convergence condition holds:* $\forall \; \theta \in \Theta$,

$$\lim_{\epsilon \to 0} \sup_{z \in \mathbb{R}} \nu(A_{z,\theta}^{\epsilon}) = 0,$$

where ν denotes the Lebesgue measure on $\mathbb{R}^n$.

Condition A.3. *The following integrability conditions hold:* $\forall \; x \in \mathbb{R}^n$, *there exist functions $v_l(\cdot), l = 1, \ldots, n$ such that $|(\partial g(x; \theta)/\partial x_i)^{-1}| \leq \prod_{l=1}^{n} v_l(x_l; \theta)$, and*

$$\lim_{x_i \to \pm\infty} v_i(x_i; \theta) f_i(x_i; \theta) = 0,$$
$$\int_{\mathbb{R}} v_l(x_l; \theta) f_l(x_l; \theta) \, dx_l < \infty, \quad l \neq i;$$

in addition,

$$\int_{x \in \mathbb{R}^n} |\Psi_{1,i}(x; \theta)| f(x; \theta) \, dx < \infty.$$

Remark 1. For distributions not supported on the whole space, for example, the exponential distribution, the continuity condition on f might not hold on the whole space, which is required in condition (A.1). However, a change of variables may be applied to transform the support to $\mathbb{R}^n$ (see Peng et al. 2018) so

that the continuity condition holds on the whole space. If $\partial g(x; \theta)/\partial x_i = 0$, then $|(\partial g(x; \theta)/\partial x_i)^{-1}|$ is interpreted as infinity. Integrability condition (A.3) implies $\partial g(x; \theta)/\partial x_i \neq 0$ a.e. For the special case when g is invertible, that is, there exists $i = 1, \ldots, n$,

$$x_i = g^{-1}(z; x_{-i}, \theta),$$

where g^{-1} means an inversion of g with respect to the ith argument and $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, condition (A.2) holds if g^{-1} is globally Lipchitz continuous with respect to z, which can be justified if there exists $\epsilon > 0$ such that $|(\partial g(x; \theta)/\partial x_i)^{-1}| > \epsilon$. With this condition, Condition (A.3) can be simplified by

$$E\left[\left\|\frac{\partial^2 g(x; \theta)}{\partial x_i^2}\right|_{x=X}\right\|\right] < \infty,$$

and

$$\lim_{x_i \to \pm\infty} f_i(x_i; \theta) = 0, \quad f(x; \theta) < \infty .$$

When g is a linear function of x, $\partial^2 g(x; \theta)/\partial x_i^2$ is zero, so the moment condition holds. The condition on the density holds for most distributions supported on the whole space.

To state our first result, we define

$$\Psi_{1,i}(x; \theta) \doteq \left(\frac{\partial g(x; \theta)}{\partial x_i}\right)^{-1}$$
$$\times \left(\frac{\partial \log f(x; \theta)}{\partial x_i} - \frac{\partial^2 g(x; \theta)}{\partial x_i^2} \left(\frac{\partial g(x; \theta)}{\partial x_i}\right)^{-1}\right).$$

We have the following theorem:

Theorem 1. *For $i = 1, \ldots, n$, under Conditions (A.1)–(A.3),*

$$\frac{\partial E[\mathbf{1}\{g(X; \theta) \leq z\}]}{\partial z} = E[\mathbf{1}\{g(X; \theta) \leq z\} \, \Psi_{1,i}(X; \theta)].$$

Proof. Define

$$\chi_\epsilon(z) \doteq \begin{cases} 1 & z < -\epsilon, \\ 1 - \frac{(z+\epsilon)}{2\epsilon} & -\epsilon \leq z \leq \epsilon, \\ 0 & z > \epsilon. \end{cases}$$

By construction, χ_ϵ is continuous and piecewise continuously differentiable. Applying the dominated convergence theorem to interchange the derivative and expectation (Fu 2006) then yields

$$\frac{\partial}{\partial z} E[\chi_\epsilon(g(X; \theta) - z)] = -E[\chi_\epsilon'(g(X; \theta) - z)]$$
$$= -\int_{x \in \mathbb{R}^n} \chi_\epsilon'(g(x; \theta) - z) f(x; \theta) \, dx,$$

where $\chi_\epsilon'(z)$ is the derivative of $\chi_\epsilon(z)$ with respect to z. We now expand the right-hand side by inserting partial derivatives of g, where we may freely choose

the particular component x_i with respect to which we differentiate. This gives

$$
\frac{\partial}{\partial z} \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\big] = -\int_{x\in\mathbb{R}^n} \chi'_\epsilon\big(g(x;\theta) - z\big) \frac{\partial g(x;\theta)}{\partial x_i} \\
\times \left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} f(x;\theta)\, dx. \quad (9)
$$

Recall that $f(x,\theta)$ has a product form, and we can write this as

$$
-\int_{\mathbb{R}^{n-1}} \left(\int_{\mathbb{R}} \chi'_\epsilon\big(g(x;\theta) - z\big) \frac{\partial g(x;\theta)}{\partial x_i}\right. \\
\left. \times \left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} f_i(x_i;\theta)\, dx_i\right) \prod_{l\neq i} f_l(x_l;\theta)\, dx_l. \quad (10)
$$

Note that

$$
\frac{\partial \chi_\epsilon\big(g(x;\theta) - z\big)}{\partial x_i} = \chi'_\epsilon\big(g(x;\theta) - z\big) \frac{\partial g(x;\theta)}{\partial x_i}.
$$

Applying integration by parts to the term in the inner bracket of (10), we obtain that (10) is equal to

$$
-\int_{\mathbb{R}^{n-1}} \chi_\epsilon\big(g(x;\theta) - z\big)\left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} f_i(x_i;\theta) \\
\times \prod_{l\neq i} f_l(x_l;\theta)dx_l \Bigg|_{x_i=-\infty}^{\infty} \\
+ \int_{\mathbb{R}^{n-1}} \left\{\int_{\mathbb{R}} \chi_\epsilon\big(g(x;\theta) - z\big) \frac{\partial}{\partial x_i}\left[\left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1}\right.\right. \\
\left.\left. \times f_i(x_i;\theta)\right] dx_i\right\} \prod_{l\neq i} f_l(x;\theta)\, dx_l \\
= \int_{\mathbb{R}^n} \chi_\epsilon\big(g(x;\theta) - z\big)\left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} \\
\times \left[\frac{\partial f(x;\theta)}{\partial x_i} - \frac{\partial^2 g(x;\theta)}{\partial x_i^2}\left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} f(x;\theta)\right] dx \\
= \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\Psi_{1,i}(X;\theta)\big],
$$

where the first equality holds because $\frac{\partial f}{\partial x_i} = \frac{\partial f_i}{\partial x_i} \prod_{l\neq i} f_l$ and by Conditions (A.1) and (A.3) and the dominated convergence theorem,

$$
\left|\int_{\mathbb{R}^{n-1}} \chi_\epsilon\big(g(x;\theta) - z\big)\left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} f_i(x_i;\theta)\right. \\
\left. \times \prod_{l\neq i} f_l(x_l;\theta)dx_l\Bigg|_{x_i=-\infty}^{\infty}\right| \\
\leq v_i(x_i;\theta) f_i(x_i;\theta)\big|_{x_i=-\infty}^{\infty} \prod_{l\neq i}\left(\int_{\mathbb{R}} v_l(x_l;\theta) f_l(x_l;\theta)\, dx_l\right) = 0.
$$

We have, thus, shown that

$$
\frac{\partial}{\partial z} \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\big] = \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\Psi_{1,i}(X;\theta)\big].
$$
(11)

It remains to be shown that taking limit $\epsilon \to 0$ yields the claim. We start off by noting that

$$
\big|\mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\Psi_{1,i}(X;\theta)\big] - \mathrm{E}\big[\mathbf{1}\{g(X;\theta) \leq z\}\Psi_{1,i}(X;\theta)\big]\big| \\
\leq \mathrm{E}\big[\mathbf{1}\{z - \epsilon \leq g(X;\theta) \leq z + \epsilon\}\big|\Psi_{1,i}(X;\theta)\big|\big] \\
= \int_{A_{z,\theta}^\epsilon} \big|\Psi_{1,i}(x;\theta)\big| f(x;\theta)\, dx.
$$

By the absolute continuity of the Lebesgue integral (Royden 1988), $\forall\, \varepsilon > 0$, $\exists\, \delta > 0$ such that, if $\nu(A_{z,\theta}^\epsilon) < \delta$, then

$$
\int_{A_{z,\theta}^\epsilon} \big|\Psi_{1,i}(x;\theta)\big| f(x;\theta)\, dx < \varepsilon.
$$

By Condition (A.2), for $\delta > 0$, $\exists\, \epsilon_0 > 0$ such that $\forall\, z \in \mathbb{R}$ and $\epsilon < \epsilon_0$, $\nu(A_{z,\theta}^\epsilon) < \delta$. Therefore,

$$
\limsup_{\epsilon\to 0}\sup_{z\in\mathbb{R}}\big|\mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\Psi_{1,i}(X;\theta)\big] \\
- \mathrm{E}\big[\mathbf{1}\{g(X;\theta) \leq z\}\Psi_{1,i}(X;\theta)\big]\big| \\
\leq \limsup_{\epsilon\to 0}\sup_{z\in\mathbb{R}} \mathrm{E}\big[\mathbf{1}\{z - \epsilon \leq g(X;\theta) \leq z + \epsilon\}\big|\Psi_{1,i}(X;\theta)\big|\big] \\
\leq \limsup_{\epsilon\to 0}\sup_{z\in\mathbb{R}} \int_{A_{z,\theta}^\epsilon} \big|\Psi_{1,i}(x;\theta)\big| f(x;\theta)\, dx = 0,
$$

which justifies the interchange of limit and derivative as follows (Rudin 1964):

$$
\frac{\partial\, \mathrm{E}\big[\mathbf{1}\{g(X;\theta) \leq z\}\big]}{\partial z} = \frac{\partial}{\partial z}\lim_{\epsilon\to 0} \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\big] \\
= \lim_{\epsilon\to 0}\frac{\partial}{\partial z} \mathrm{E}\big[\chi_\epsilon\big(g(X;\theta) - z\big)\big] \\
= \mathrm{E}\big[\mathbf{1}\{g(X;\theta) \leq z\}\, \Psi_{1,i}(X;\theta)\big],
$$

and proves the claim. $\quad\square$

Remark 2. The GLR estimator in Theorem 1 is not unique. Different weight functions $\Psi_{1,i}$, $i = 1,\ldots,n$, correspond to the integration by parts with respect to different coordinates of input r.v. $X = (X_1,\ldots,X_n)$. An alternative interpretation of the GLR lies in the differentiation of an implicit change of variable (see the online appendix in Peng et al. 2018). Different weight functions also correspond to the changing variable in different coordinates. Within an unbiased GLR estimator family for the distribution sensitivities, we can obtain an optimal estimator by minimizing variance; see Section A.3 in the online appendix.

Let $\Psi_{j,i}$ be defined in a jth-order distribution sensitivity estimator such that

$$\frac{\partial^j \mathrm{E}[\mathbf{1}\{g(X;\theta) \le z\}]}{\partial \theta^{j-1}\partial z} = \mathrm{E}[\mathbf{1}\{g(X;\theta) \le z\}\ \Psi_{j,i}(X;\theta)].$$

To derive the GLR estimators for the $(j+1)$th order distribution sensitivities, we introduce the following regularity condition.

Condition A.4. *The following uniform convergence condition holds:* $\forall\ z \in \mathbb{R}$,

$$\lim_{\epsilon \to 0} \sup_{\theta \in \Theta} \nu(A_{z,\theta}^{\epsilon}) = 0.$$

Condition A.5. *The following integrability conditions hold: for any $j \in \mathbb{Z}^+$,*

$$\lim_{x_i \to \pm\infty} \int_{\mathbb{R}^{n-1}} \left| \left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} \frac{\partial g(x;\theta)}{\partial \theta} \Psi_{j,i}(x;\theta) \right| f(x;\theta)$$

$$\times \prod_{l \ne i} dx_l = 0,$$

and

$$\int_{x \in \mathbb{R}^n} \sup_{\theta \in \Theta} \left| \Psi_{j+1,i}(x;\theta)\, f(x;\theta) \right| dx < \infty,$$

where

$$\Psi_{j+1,i}(x;s,\theta) \doteq \frac{\partial \log f(x;\theta)}{\partial \theta} + \frac{\partial \Psi_{j,i}(x;\theta)}{\partial \theta}$$
$$- \left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} \left\{ \frac{\partial^2 g(x;\theta)}{\partial\theta\partial x_i} \Psi_{j,i}(x;\theta) \right.$$
$$+ \frac{\partial g(x;\theta)}{\partial \theta} \left[\frac{\partial \Psi_{j,i}(x;\theta)}{\partial x_i} + \Psi_{j,i}(x;\theta) \right.$$
$$\left. \left. \times \left(\frac{\partial \log f(x;\theta)}{\partial x_i} - \frac{\partial^2 g(x;\theta)}{\partial x_i^2} \left(\frac{\partial g(x;\theta)}{\partial x_i}\right)^{-1} \right) \right] \right\}.$$

Note that the quantity $\Psi_{j+1,i}(x;s,\theta)$ is defined recursively in terms of its lower-order analog $\Psi_{j,i}(x;s,\theta)$.

Theorem 2. *For $i = 1,\ldots,n$, under Conditions* (A.1), (A.2), (A.4), *and* (A.5) *for any $j \in \mathbb{Z}^+$,*

$$\frac{\partial^{j+1} \mathrm{E}[\mathbf{1}\{g(X;\theta) \le z\}]}{\partial \theta^j \partial z} = \mathrm{E}[\mathbf{1}\{g(X;\theta) \le z\}\ \Psi_{j+1,i}(X;\theta)].$$

The proof of Theorem 2 is similar to that of Theorem 1 and can be found in Section A.1 of the online appendix. Note that the implementation of the GLR estimators for distributional sensitivities requires the capability to simulate the input r.v. X. The convergence rates of these estimators in terms of the simulation replication size are canonical square root and, moreover, with a bound on the first-order multiplicative constant that is independent of the realization of

the real-world observations. This is justified by a uniform convergence of the GLR estimators over z (see Section A.2 of the online appendix for details). Moreover, note that, for estimating the density and the derivatives for multiple parameters, one can use the same batch of simulation samples and apply the corresponding $\Psi_{j,i}(X;\theta)$ for each of these estimators.

3.2. Sample Path Derivative for HMM

Assuming the observation kernel p and its derivatives known (or can be estimated) in HMM, we derive an IPA estimator for the derivative of the log-likelihood (5). To facilitate calculation, we assume $Y_{i,t}$, $i = 1,\ldots,n$ are independent, $t = 1,\ldots,T$, with marginal distribution function and density given by $Q_i(\cdot;\theta)$ and $q_i(\cdot;\theta)$, respectively. Assuming the derivative and expectation can be interchanged, which is typically justified by the dominated convergence theorem (Glasserman 1991),

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\mathrm{E}\left[\prod_{t=1}^T p(Z_t; S_t, \theta)\right] \Big/ \mathrm{E}\left[\prod_{t=1}^T p(Z_t; S_t, \theta)\right]$$

$$= \mathrm{E}\left[\left(\sum_{t=1}^T W_t(S_t;\theta)\right)\prod_{t=1}^T p(Z_t; S_t, \theta)\right] \Big/$$

$$\times \mathrm{E}\left[\prod_{t=1}^T p(Z_t; S_t, \theta)\right], \tag{12}$$

where

$$W_t(S_t;\theta) \doteq \left(\frac{\partial p(Z_t; S_t, \theta)}{\partial \theta} + \frac{\partial p(Z_t; s, \theta)}{\partial s}\bigg|_{s=S_t} \frac{\partial S_t}{\partial \theta}\right)$$

$$\times \left(p(Z_t; S_t, \theta)\right)^{-1}, \tag{13}$$

and for $t = 1,\ldots,T$,

$$\frac{\partial S_t}{\partial \theta} = \frac{\partial h(Y_t; S_{t-1};\theta)}{\partial \theta} + \sum_{i=1}^n \frac{\partial h(y; S_{t-1};\theta)}{\partial y_i}\bigg|_{y=Y_t} \frac{\partial Y_{i,t}}{\partial \theta}$$

$$+ \frac{\partial h(Y_t; s;\theta)}{\partial s}\bigg|_{s=S_{t-1}} \frac{\partial S_{t-1}}{\partial \theta},$$

$$\frac{\partial Y_{i,t}}{\partial \theta} = -\frac{\partial Q_i(Y_{i,t};\theta)}{\partial \theta}\bigg/ q_i(Y_{i,t};\theta), \tag{14}$$

where $y \doteq (y_1,\ldots,y_n)$. Note that the expression for the derivative of $Y_{i,t}$ is an IPA estimator; see Suri and Zazanis (1988).

In our setting, the conditional density of the observation on the hidden state may not be analytically known. The LR estimator for the derivative of log-likelihood (5) can also be derived, but it requires estimating the conditional density of the hidden Markov chain and its derivatives besides estimating the observational kernel and associated derivatives, which adds extra computational burden.

For more general HMM (4), the sensitivity of the density with respect to the hidden state variable S_t might also be needed. Similar to the proof of Theorem 2, under appropriate regularity conditions, we have the distribution sensitivity w.r.t. the hidden state variable s and the argument z:

$$\frac{\partial^2 \, \mathrm{E}\big[\mathbf{1}\{g(X;s,\theta) \le z\}\big]}{\partial s \partial z} = \mathrm{E}\Big[\mathbf{1}\{g(X;s,\theta) \le z\}\, \widetilde{\Psi}_{2,i}(X;s,\theta)\Big],$$

where

$$
\begin{aligned}
&\widetilde{\Psi}_{2,i}(x;s,\theta)\\
&\doteq \frac{\partial \Psi_{1,i}(x;s,\theta)}{\partial s} - \left(\frac{\partial g(x;s,\theta)}{\partial x_i}\right)^{-1}\Bigg\{\frac{\partial^2 g(x;s,\theta)}{\partial s \partial x_i}\Psi_{1,i}(x;s,\theta)\\
&\quad + \frac{\partial g(x;s,\theta)}{\partial s}\Bigg[\frac{\partial \Psi_{1,i}(x;s,\theta)}{\partial x_i} + \Psi_{1,i}(x;s,\theta)\\
&\quad \times \left(\frac{\partial \log f(x;\theta)}{\partial x_i} - \frac{\partial^2 g(x;s,\theta)}{\partial x_i^2}\left(\frac{\partial g(x;s,\theta)}{\partial x_i}\right)^{-1}\right)\Bigg]\Bigg\}.
\end{aligned}
$$

Simulation is needed to implement the derivative estimator given by (12) for the HMM. Sampling from the prior distribution of hidden Markov chain is straightforward, but the variance of the estimate may be extremely high if the prior distribution and posterior distribution differ significantly, which is very likely if the number of observations is large.

To overcome the drawbacks of direct sampling, we sample from the posterior distribution, which is a consecutive update of the prior distribution by incorporating information from observations sequentially, using the Bayes rule. This is often known as the filtering problem. Because of the sequential structure of the HMM, sequential Monte Carlo (SMC) can be used as an iterative sampling algorithm of the filtering measure, which provides estimators of the conditional expectation with relatively lower variance. Because of page limitations, we introduce only the minimal background on Bayesian statistics and SMC sufficient for understanding our proposed method. For more details, see Doucet (2001).

By the Bayes rule, the posterior density of the hidden states $S_{1:t+1} \doteq (S_1, \ldots, S_{t+1})$ conditional on $Z_{1:t} \doteq (Z_1, \ldots, Z_t)$ is

$$\pi_{t+1|t}(d \,\cdot) \doteq \frac{\prod_{\ell=1}^{t} p(Z_\ell; S_\ell, \theta)\, p(S_{\ell+1}; S_\ell, \theta)\, d \,\cdot}{\int_{\mathbb{R}^{t+1}} \prod_{\ell=1}^{t} p(Z_\ell; S_\ell, \theta)\, p(S_{\ell+1}; S_\ell, \theta)\, dS_{1:t+1}},$$

where $\prod_{\ell=1}^{0} \equiv 1$. Thus, the conditional expectation of $p_{t+1}(S_{t+1}; \theta)$ is

$$\pi_{t+1|t}(p_{t+1}) = \mathrm{E}\left[\prod_{\ell=1}^{t+1} p_\ell(S_\ell;\theta)\right]\Bigg/ \mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right], \quad (15)$$

where $p_t(S_t;\theta) \doteq p(Z_t; S_t, \theta)$. To apply SMC, we decompose the log-likelihood (5) into a sum of log conditional expectations

$$L_T(\theta) = \sum_{t=0}^{T-1} \log \pi_{t+1|t}(p_{t+1}) \, .$$

Taking the derivative, we have

$$\frac{\partial L_T(\theta)}{\partial \theta} = \sum_{t=0}^{T-1} \frac{\partial \pi_{t+1|t}(p_{t+1})}{\partial \theta}\Bigg/ \pi_{t+1|t}(p_{t+1}),$$

and elaborating on (15), we obtain

$$
\begin{aligned}
\frac{\partial \pi_{t+1|t}(p_{t+1})}{\partial \theta} &= \frac{\partial}{\partial \theta}\mathrm{E}\left[\prod_{\ell=1}^{t+1} p_\ell(S_\ell;\theta)\right]\Bigg/ \mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right]\\
&\quad - \pi_{t+1|t}(p_{t+1})\frac{\partial}{\partial \theta}\mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right]\Bigg/\\
&\qquad \mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right]\\
&= \mathrm{E}\left[\left(\sum_{\ell=1}^{t+1} W_\ell(S_\ell;\theta)\right)\prod_{\ell=1}^{t+1} p_\ell(S_\ell;\theta)\right]\Bigg/\\
&\qquad \mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right]\\
&\quad - \pi_{t+1|t}(p_{t+1})\,\mathrm{E}\left[\left(\sum_{\ell=1}^{t} W_\ell(S_\ell;\theta)\right)\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right]\\
&\qquad \Bigg/ \mathrm{E}\left[\prod_{\ell=1}^{t} p_\ell(S_\ell;\theta)\right],
\end{aligned}
$$

with $W_\ell(S_\ell;\theta)$ defined in (13).

If we can sample from the posterior distribution $\pi_{t+1|t}(\cdot)$ of the hidden states $S_{1:t+1}$ given observations $Z_{1:t}$, then the posterior expectation $\pi_{t+1|t}(\varphi)$ for a measurable function $\varphi(\cdot)$ on $\mathbb{R}^{t+1}$ has the following unbiased estimator:

$$\widetilde{\pi}_{t+1|t}^{J}(\varphi) \doteq \frac{1}{J}\sum_{j=1}^{J}\varphi\left(\widetilde{S}_{1:t+1}^{(j)}\right),$$

where $\widetilde{S}_{1:t+1}^{(j)}$ is the jth particle (sample path) of the hidden state generated from the posterior distribution. However, it is infeasible to directly sample from the posterior distribution because there is no closed form for the high-dimensional integral in the posterior distribution. Alternatively, we can sample from the prior distribution (Markov chain) of $S_{1:t+1}$ and then use importance sampling to reweight each sample in the estimator of the posterior expectation as follows:

$$\pi_{t+1|t}^{J}(\varphi) \doteq \frac{1}{J}\sum_{j=1}^{J}\varphi\left(S_{1:t+1}^{(j)}\right)w_{1:t+1}^{(j)},$$

where $S_{1:t+1}^{(j)}$ is the jth particle of the hidden state generated from the prior distribution, and

$$w_{1:t+1}^{(j)} \doteq \frac{\prod_{\ell=1}^{t} p_\ell\left(S_\ell^{(j)}; \theta\right)}{\sum_{j=1}^{J} \prod_{\ell=1}^{t} p_\ell\left(S_\ell^{(j)}; \theta\right)}.$$

However, estimator $\pi_{t+1|t}^{J}(\varphi)$ might end up with extremely large variance. An intuitive explanation is that the prior distribution of $S_{1:t+1}$, following a long Markov chain, might differ significantly from the posterior distribution adjusted by observations, so a large portion of particles would have their weights ($w_{1:t+1}^{(j)}$, $j = 1, \ldots, J$) close to zero.

In SMC, the particles are propagated over time using a combination of sequential importance sampling and resampling steps. The resampling step statistically multiplies and discards particles at each step to adaptively concentrate particles on the region of high intensity of the posterior distribution. Specifically, SMC updates the posterior distribution by the following sequential mechanism:

$$\widehat{\pi}_{\ell|\ell}^{J}(\cdot) \doteq \frac{1}{J} \sum_{j=1}^{J} \delta_{\widehat{S}_{1:\ell}^{(j)}}(\cdot) \to \widehat{\pi}_{\ell+1|\ell}^{J}(\cdot)$$

$$\doteq \frac{1}{J} \sum_{j=1}^{J} \delta_{\left(\widehat{S}_{1:\ell}^{(j)}, S_{\ell+1}^{(j)}\right)}(\cdot) \to \widehat{\pi}_{\ell+1|\ell+1}^{J}(\cdot)$$

$$\doteq \frac{1}{J} \sum_{j=1}^{J} \delta_{\widehat{S}_{1:\ell+1}^{(j)}}(\cdot),$$

where $\delta_{\widehat{S}_{1:\ell}^{(j)}}(\cdot)$ is a δ-measure concentrated on $\widehat{S}_{1:\ell}^{(j)}$, $S_{\ell+1}^{(j)}$ is sampled from the transition function of the hidden Markov chain:

$$S_{\ell+1}^{(j)} = h\left(Y_\ell^{(j)}; \widehat{S}_\ell^{(j)}, \theta\right),$$

and $\widehat{S}_{\ell+1}^{(j)}$, $j = 1, \ldots, J$, are resampled from $S_{\ell+1}^{(j)}$, $j = 1, \ldots, J$ with weights adjusted by the $\ell + 1$th observation

$$\widehat{w}_{\ell+1}^{(j)} \doteq \frac{p_{\ell+1}\left(S_{\ell+1}^{(j)}; \theta\right)}{\sum_{j=1}^{J} p_{\ell+1}\left(S_{\ell+1}^{(j)}; \theta\right)}, \quad j = 1, \ldots, J.$$

SMC is consistent and can significantly reduce the variance of the estimator but introduces a slight bias that decreases linearly with the number of particles. Under appropriate regularity conditions,

$$\lim_{J \to \infty} \widehat{\pi}_{t+1|t}^{J}(\varphi) = \pi_{t+1|t}(\varphi) \quad \text{a.s.},$$

$$\mathrm{E}\left[\widehat{\pi}_{t+1|t}^{J}(\varphi)\right] = \pi_{t+1|t}(\varphi) + O(J^{-1}).$$

In addition, its (asymptotic) variance is $O(J^{-1})$. The proofs for the asymptotic variance and bias estimate can be found in Del Moral (2004, chapters 8 and 9).

Generating r.v.s Y_t, $t = 1, \ldots, T$, driving the hidden Markov chain might be time consuming in practical applications (Peng et al. 2014, 2016). We can use one batch of i.i.d. simulated samples $Y^{(i)}$, $i = 1, \ldots, N$, of the i.i.d. r.v.s Y_t, $t = 1, \ldots, T$, driving the hidden Markov chain in (4). For $t = 1, \ldots, T$, $Y_t^{(j)}$ is resampled independently from $\{Y^{(i)}\}_{i=1}^{N}$, $j = 1, \ldots, J$. To achieve this, we only need to generate T independent multinomial r.v.s, which is computationally cheap, to obtain T i.i.d. copies of r.v.s generated from the empirical distribution:

$$\mathbb{Q}_N(\cdot) = \frac{1}{N} \sum_{j=1}^{N} \delta_{Y^{(j)}}(\cdot).$$

We define the expectation with respect to the empirical distribution as follows:

$$\mathbb{E}_N\left[\varphi(Y_1, \ldots, Y_t)\right] \doteq \int_{\mathbb{R}^t} \varphi(y_1, \ldots, y_t) \prod_{\ell=1}^{t} \mathbb{Q}_N(dy_\ell).$$

To avoid generating input r.v.s in every iteration of SA, we apply a change of measure:

$$\mathrm{E}\left[\varphi(Y_1, .., Y_t)\right] = \mathrm{E}\left[\varphi(\bar{Y}_1, .., \bar{Y}_t) \prod_{\ell=1}^{t} R(\bar{Y}_\ell; \theta, \bar{\theta})\right],$$

where $\bar{Y}_\ell$, $\ell = 1, \ldots, t$, are i.i.d. random vectors with density $q(\cdot; \bar{\theta})$ and

$$R(x; \theta, \bar{\theta}) \doteq q(x; \theta)/q(x; \bar{\theta}). \quad (16)$$

With parameter θ updating in SA, the distribution of input r.v.s under θ could get further away from the distribution of input r.v.s under $\bar{\theta}$. The degeneracy of samples can be measured by the effective sample size (ESS) criterion (Liu and Chen 1998)

$$ESS \doteq \left\{\sum_{j=1}^{N} \left(\bar{w}^{(j)}\right)^2\right\}^{-1},$$

where $\bar{w}^{(j)} \doteq R^{(j)}/(\sum_{i=1}^{N} R^{(i)})$ and $R^{(j)} = R(\bar{Y}^{(j)}; \bar{\theta}, \theta)$, which take values between one and M. If the degeneracy is too high, that is, ESS is below a prespecified threshold, say $M/3$, then we resimulate a new batch of input r.v.s.

What remains to be addressed is the fact that the product of empirical distributions $\prod_{\ell=1}^{t} \mathbb{Q}_N(dy_\ell)$ is obtained by one batch of simulated samples instead of t batches of independently generated samples, so the law of large numbers does not apply. In Section A.4 in the online appendix, in which the theory of empirical processes is elaborated, we establish consistency and

a central limit theorem for the empirical expectation $\mathbb{E}_N[\varphi(Y_1,..,Y_t)]$.

4. Applications

In this section, we apply GSMLE to the i.i.d. case in Section 4.1, to the Markovian model in Section 4.2, and to the HMM in Section 4.3. The algorithms for implementation can be found in Section A.6 of the online appendix, and the codes can also be found in the online supplemental material.

4.1. I.I.D. Case

We suppose i.i.d. observations are generated by a data-generating process $g(X_t; \theta) = X_{1,t} + \theta X_{2,t}$, where $X_{1,t}, X_{2,t} \sim N(0,1)$ are independent. For this example, the MLE has an analytical form:

$$\hat{\theta} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} Z_t^2 - 1}.$$

From Theorems 1 and 2, the two GLR estimators with $i = 1$ for the distributional sensitivities in (7), using a single simulation run, are

$$-\mathbf{1}\{g(X_t;\theta) \le z\}X_{1,t}, \quad \mathbf{1}\{g(X_t;\theta) \le z\}X_{2,t}(1 - X_{1,t}^2).$$

From Remark 2 after Theorem 1, there are different choices of the weight functions in the GLR estimators. Here we choose a GLR estimator that has the simplest weight function, namely putting all weight on $i = 1$. In Online Appendix A.3, we derive a variance-minimizing estimator and numerically compare different choices of the weight function. We use simulated samples of $X_t = (X_{1,t}, X_{2,t})$ to estimate the likelihood function and its derivative. The true value is set to $\theta = 1$. The step size in SA algorithm (6) is chosen as $\lambda_k = a/k$ with $a = 0.01$, starting point $\theta_0 = 0.8$, and feasible set $\Theta = [0.5, 2]$. We take M, the number of simulated samples per iteration, to be 10^5 or 10^4 in our implementation. Such a large number of samples is used to ensure a negligible bias coming from the ratio form of our log-likelihood derivative estimator. In fact, because θ does not show up in the distribution of X_t and we have used a large simulated sample size, we opt to reuse these samples across iterations (i.e., one can generate new samples in each iteration, but this is unlikely to substantially affect the subsequent observed algorithmic behaviors). A similar approach is applied to other examples in this section.

To see the statistical behavior of SA algorithm (6) with simulated derivative estimates for the likelihood of a fixed set of observations, we generate one batch of $T = 100$ i.i.d. observations. In Figure 1, the GSMLE converges in about 50 iterations. The true MLE (black line) lies within the confidence interval of the GSMLE

with $M = 10^4$ simulated samples (blue line) and $M = 10^5$ simulation samples (red line) based on 100 independent experiments. The GSMLE with $M = 10^5$ simulated samples has smaller bias and standard error than the GSMLE with $M = 10^4$ simulated samples.

To test the statistical behavior of the GSMLE, we independently generate 100 batches of 100 and 1,000 i.i.d. observations and report estimation results in boxplots. The horizontal (black) line in the middle is the true value. The (red) line in the middle of each box is the sample median. The tops and bottoms of each box are the 25th and 75th percentiles of the estimates, respectively. From Figure 2, we can see the average behavior of the GSMLE with $M = 10^5$ simulated samples is similar to the statistical behavior of the true MLE.

If we assume $X_{1,t} \sim N(0,1)$ and $X_{2,t} \sim t_2$, where t_2 is a t-distribution with two degrees of freedom, the MLE does not have an analytical form. The GLR estimators for the distribution sensitivities can be obtained similarly, and for $i = 1$ in Theorems 1 and 2, the GLR estimators have the same form as in the case in which $X_{2,t}$ is a standard normal r.v. We run GSMLE with $M = 10^6$, $M = 10^7$, and $M = 10^8$ simulated samples under the same initialization of the algorithm as the previous case. In Figure 3, we report the boxplots of GSMLE based on independent 100 batches of 100 and 1,000 i.i.d. observations. We can see that it requires more simulated samples to reduce the ratio bias in calculating the GSMLE compared with the previous case.

4.2. Markovian Case

In this section, GSMLE is applied to both a simple Markovian example with an analytical likelihood and a queueing example without an analytical likelihood.

Figure 1. (Color online) Trajectories (Solid Lines) of the Means of the GSMLE with Sample Size $M = 10^4$ (Blue Line in Online Version) and $M = 10^5$ (Red Line in Online Version) Bounded by the Trajectories of Means ± Standard Errors (Dotted Line) Based on 100 Independent Experiments for the Linear Gaussian System
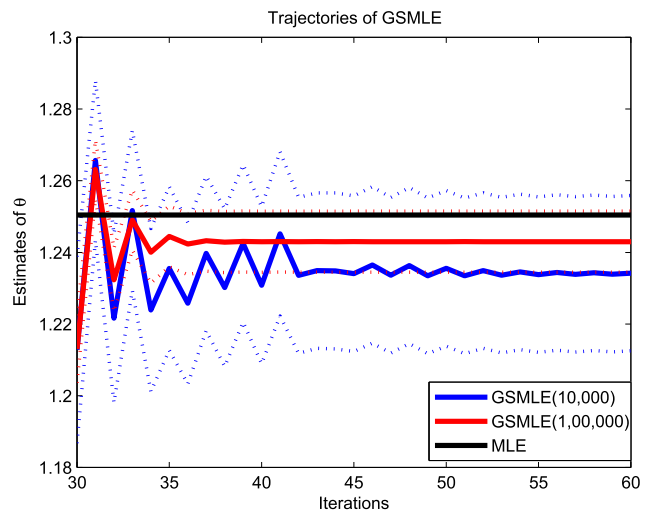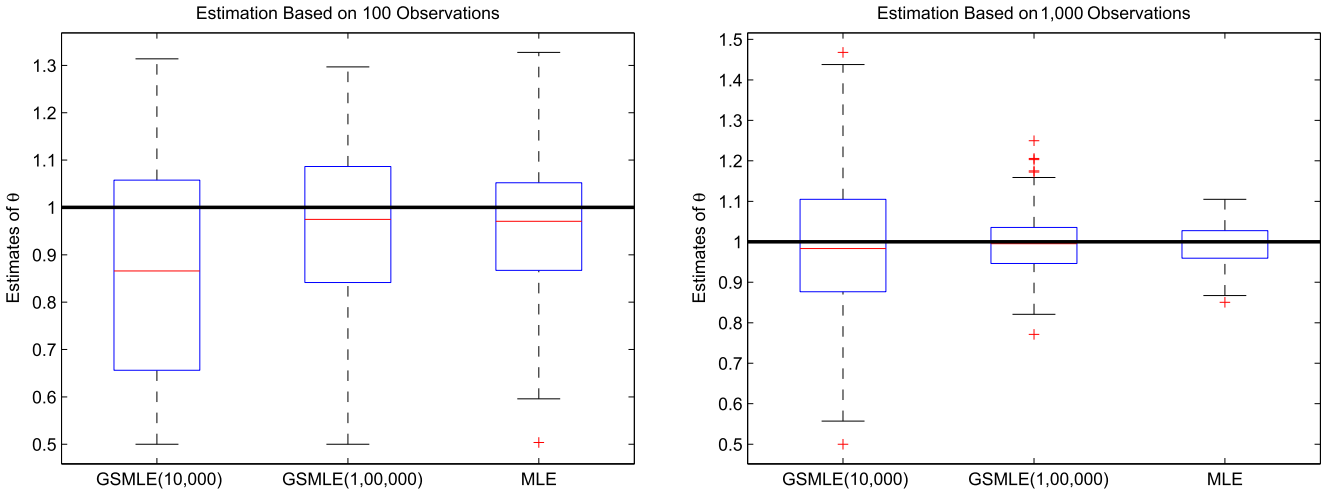
Figure 2. (Color online) Boxplots of MLE and GSMLE with 100 and 1,000 Observations Based on 100 Independent Experiments for $X_{1,t}, X_{2,t} \sim N(0,1)$

In the queueing example, we illustrate the difference between the input fitting and output fitting.

4.2.1. Simple Example. In this example, we test the performance of the GSMLE on a simple Markov process (autoregressive model of order 1) given by the following data-generating process: for $t = 1, \dots, T$,

$$Z_t = g(X_t; Z_{t-1}, \theta) \doteq \theta Z_{t-1} + X_t,$$

where X_t, $t = 1, \dots, T$, and Z_0 are i.i.d. standard normal distributed r.v.s. The necessary and sufficient condition for the time series model to have a stationary distribution is $\theta < 1$. For this example, the MLE has the following analytical form:

$$\hat{\theta} = \frac{\sum_{t=1}^{T} Z_{t-1} Z_t}{\sum_{t=1}^{T} Z_{t-1}^2} .$$

By Theorem 1, the GLR estimators for distribution sensitivities using a single simulation run are

$$-\mathbf{1}\{g(X_t; Z_{t-1}, \theta) \le z\} X_t,$$
$$\mathbf{1}\{g(X_t; Z_{t-1}, \theta) \le z\} Z_{t-1}(1 - X_t^2) .$$

We use one batch of M simulated samples of X_t to estimate the likelihood function and its derivative. The true value is set to $\theta = 0.5$. The step size in SA algorithm (6) is chosen as $\lambda_k = a/k$ with $a = 0.01$, the starting point $\theta_0 = 0.3$, and the feasible set $\Theta = [0.2, 0.8]$.

Figure 4 presents the boxplots of 100 independent GSMLEs with 100 and 1,000 observations. In this example, the GSMLE with $M = 10^4$ simulated samples, GSMLE with $M = 10^5$ simulated samples, and the true MLE have comparable statistical performances.

4.2.2. Queueing Example. We implement GSMLE on Example 2 of Section 2. Assume both $B_t(\theta)$ and A_t

Figure 3. (Color online) Boxplots of MLE and GSMLE with 100 and 1,000 Observations Based on 100 Independent Experiments for $X_{1,t} \sim N(0,1)$ and $X_{2,t} \sim t_2$
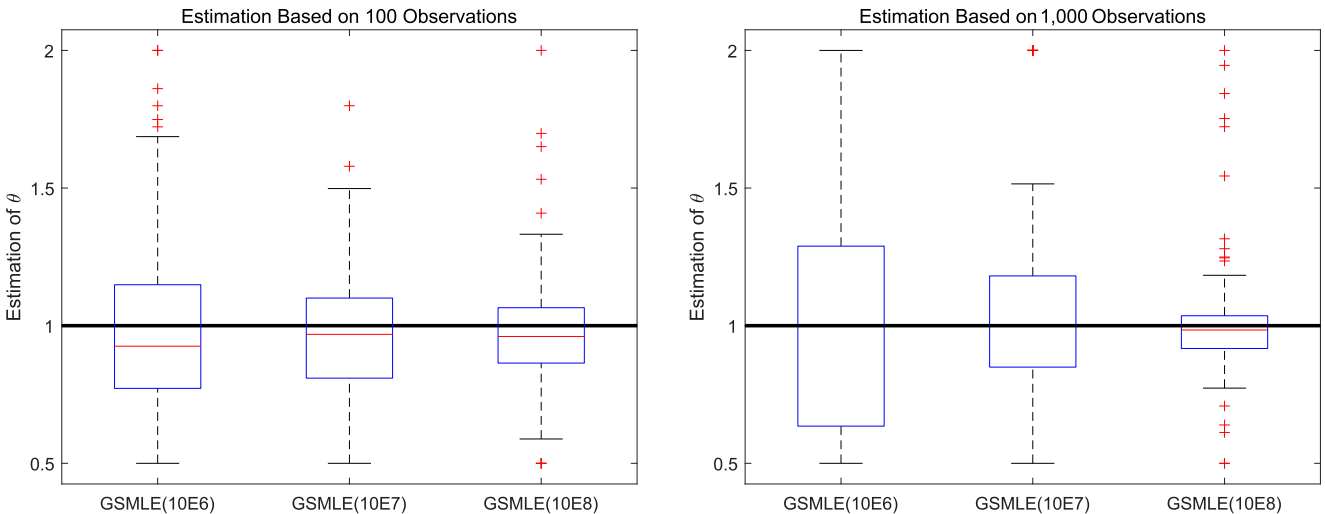
follow log-normal distributions. Let $X_{1,t}$ and $X_{2,t}$ be independent standard normal r.v.s, and $B_t(\theta) = e^{\sigma_1 X_{1,t} + \theta}$ and $A_t = e^{\sigma_2 X_{2,t} + \mu_2}$. We can rewrite the model as follows:

$$g(X_t; Z_{t-1}, \theta) = \max\{0, Z_{t-1}(\theta) - e^{\sigma_2 X_{2,t} + \mu_2}\} + e^{\sigma_1 X_{1,t} + \theta}, \quad t \geq 1.$$

From Theorems 1 and 2, the GLR estimators with $i = 1$ for distribution sensitivities are

$$-\frac{1}{\sigma_1} \mathbf{1}\{g(X_t; Z_{t-1}, \theta) \leq z\}(X_{1,t} + \sigma_1)e^{-(\theta + \sigma_1 X_{1,t})},$$

$$-\frac{1}{\sigma_1} \mathbf{1}\{g(X_t; Z_{t-1}, \theta) \leq z\}$$
$$\times \left[\frac{(X_{1,t} + \sigma_1)^2}{\sigma_1} - (X_{1,t} + \sigma_1) - \frac{1}{\sigma_1}\right]e^{-(\theta + \sigma_1 X_{1,t})}.$$

Notice that the exponential term in the estimator could lead to large variance. To further reduce the variance, we can do a change of variables. Alternative estimators can be written as

$$-\frac{e^{\sigma_1^2/2 - \theta}}{\sigma_1} \mathbf{1}\{\tilde{g}(X_t; Z_{t-1}, \theta) \leq z\}X_{1,t},$$

$$-\frac{e^{\sigma_1^2/2 - \theta}}{\sigma_1} \mathbf{1}\{\tilde{g}(X_t; Z_{t-1}, \theta) \leq z\}\left[\frac{X_{1,t}^2}{\sigma_1} - X_{1,t} - \frac{1}{\sigma_1}\right],$$

where

$$\tilde{g}(X_t; Z_{t-1}, \theta) = \max\{0, Z_{t-1}(\theta) - e^{\sigma_2 X_{2,t} + \mu_2}\} + e^{\sigma_1 X_{1,t} - \sigma_1^2 + \theta}, \quad t \geq 1.$$

Let $\mu_2 = \sigma_1 = \sigma_2 = 1$ and $\theta = 0$. Set $\lambda_k = a/k$ with $a = 0.1$, starting point $\theta_0 = 0.5$, and feasible set $\Theta = [-1, 1]$. Similar to the i.i.d. case, we show the statistical behavior for the trajectory of SA based on $T = 100$

observations and $M = 10^5, 10^6$ simulated samples. MLE-input is the MLE for θ, assuming the input random variables B_t, $t = 1, \dots, T$, are observable:

$$\frac{1}{T}\sum_{t=1}^{T} \log B_t.$$

From Figure 5, we can see that GSMLE converges very fast and is almost identical to the MLE-input when the sample size reaches 10^7.

Figure 6 presents the boxplots of 100 independent GSMLEs with 100 observations under the true model and model misspecification. In the case of model misspecification, we generate the observations from Example 3 of Section 2. Let $\zeta_1 = e^{X_{2,t}}$, $\zeta_2 = e^{2 + X_{3,t}}$, where $X_{2,t}$ and $X_{3,t}$ are independent standard normal r.v.s, and $\eta = 0.4$. So A_t switches between ζ_1 and ζ_2, following a Markov chain. On the other hand, we calculate the GSMLE based on Example 2 of Section 2 and let $A_t = e^{1 + X_{2,t}}$.

From Figure 6, we can see that MLE-input produces an estimate close to the true service time distribution parameter value ($\theta = 0$), whereas the GSMLEs that are calculated based on the likelihood of the output observations provide estimates significantly different from the true value. In contrast, GSMLE seems to perform better than MLE-input in terms of the accuracy on output performance measures. In Figure 7, we can see that the expected system times of the first 10 customers in the Lognormal/Lognormal/1 queueing model with the service rate calibrated by the MLE using the input data grows at a much slower speed than those in the true model, whereas the growth rate of the expected system times of the first six customers in the Lognormal/Lognormal/1 model with service rate calibrated by the GSMLE(10^6) using the output data catches up with the growth rate of the

Figure 4. (Color online) Boxplots of MLE and GSMLE with 100 and 1,000 Observations Based on 100 Independent Experiments for a Simple Markov Model
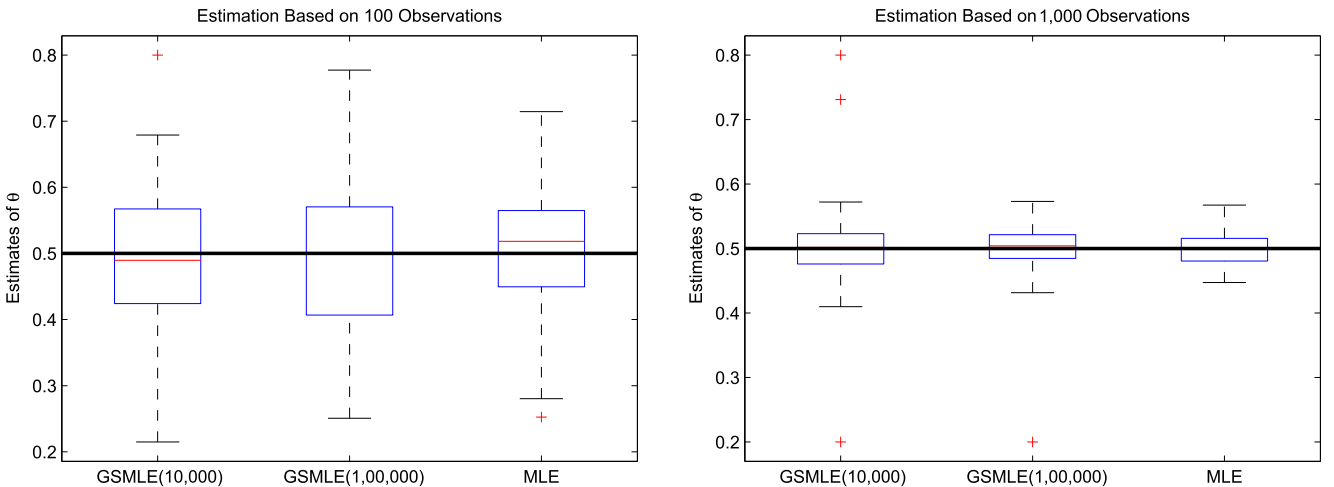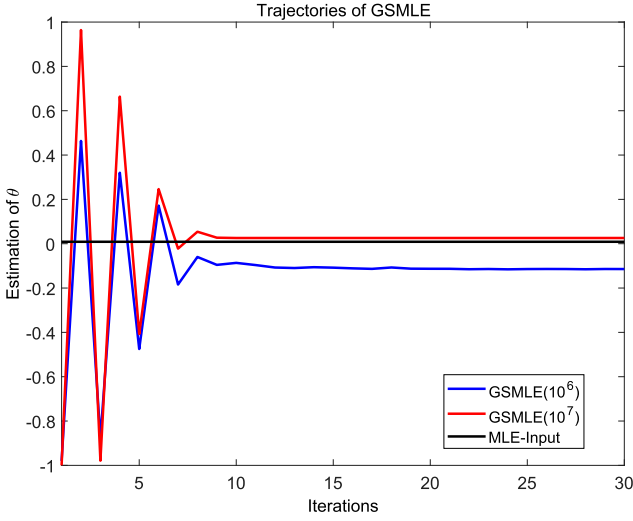
Figure 5. (Color online) Trajectories of the Means of the GSMLE with Sample Size $M = 10^6$ (Blue Line) and $M = 10^7$ (Red Line) Based on 100 Independent Experiments for the Queueing Example

expected system times of the corresponding customers in the true model though the rate lags behind after nine customers.

In Table 1, we show the estimates of the average system time of the first 10 customers with data generated by the true model, the misspecified model with the service rate fitted by GSMLE(10^6), and the misspecified model with the service rate fitted by the MLE-input. In this example, we can see the average system time of the misspecified model with the service rate fitted by the GSMLE is much closer to the average system time of the true model than the average system time of the misspecified model fitted by the MLE-input. The observation indicates that even fitting one parameter using output data can

significantly push the statistical property of a misspecified M/M/1 queueing model "closer" to the statistical property of the true model generating the data. An experiment using real bank data can be found in Section A.5 of the online appendix.

4.3. Hidden Markov Model

For the HMM, GSMLE can be implemented by Algorithm 1 provided in Online Appendix Section A.5, and Algorithm 3 outputs the derivative estimator of log-likelihood (5) by using the SMC approximation discussed in Section 3.2. The estimators in Algorithm 3 implement a basic SMC sampler to sample from the posterior distribution. We apply the same techniques used in Peng et al. (2014) to further reduce variance and enhance efficiency.

In Algorithm 1, one batch of input random vectors $\{\bar{X}^{(m)}\}_{m=1}^M$ and $\{\bar{Y}^{(j)}\}_{j=1}^N$ obtained by simulating the underlying state space model with a fixed parameter $\bar{\theta}$ is used throughout the experiment. This can significantly reduce the number of simulated samples from $K \times (T+1) \times (M+J)$ required in a straightforward simulation procedure for estimating the derivative of log-likelihood (5) to $M + N + J$.

We implement the GSMLE on Example 4 of Section 2. Assume X_t follows a standard normal distribution. From Theorems 1 and 2, the distribution sensitivities are given by

$$-\mathbf{1}\{g(X_t; S_t, \theta) \leq z\} X_t,$$
$$-\mathbf{1}\{g(X_t; S_t, \theta) \leq z\} (X_t^2 - 1)(c_1 \mathbf{1}\{S_t \leq \theta\} + c_2 \mathbf{1}\{S_t \geq \theta\}).$$

The interarrival times and service times are assumed to be lognormally distributed with mean zero and variance one and mean one and variance one in the normal distributions, respectively. We set behavior

Figure 6. (Color online) Boxplots of MLE and GSMLE with 100 Observations Under the True Model and Model Misspecification Based on 100 Independent Experiments for a Queueing Example
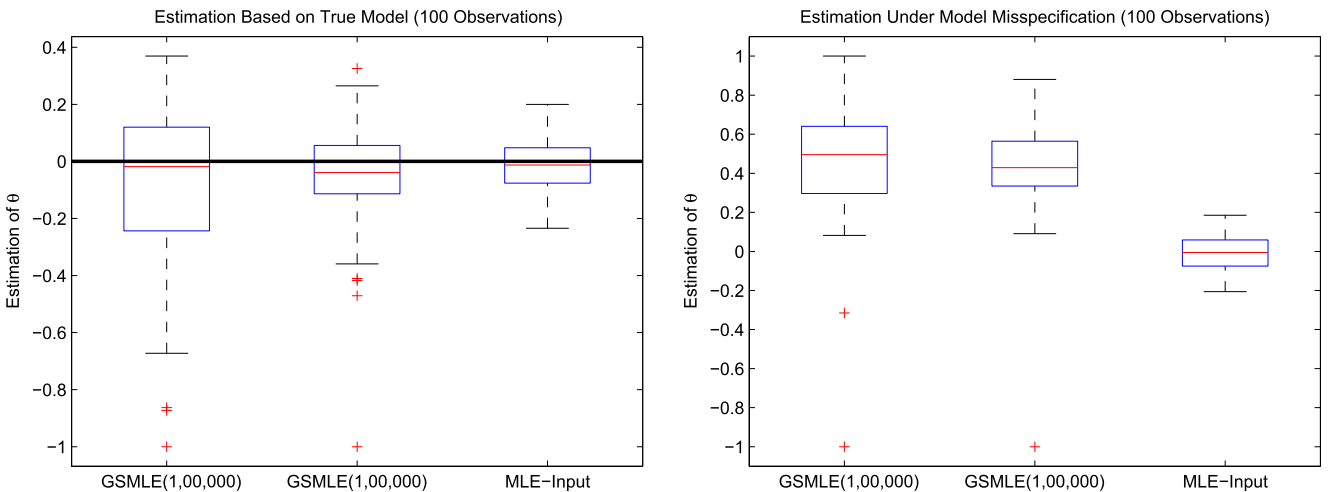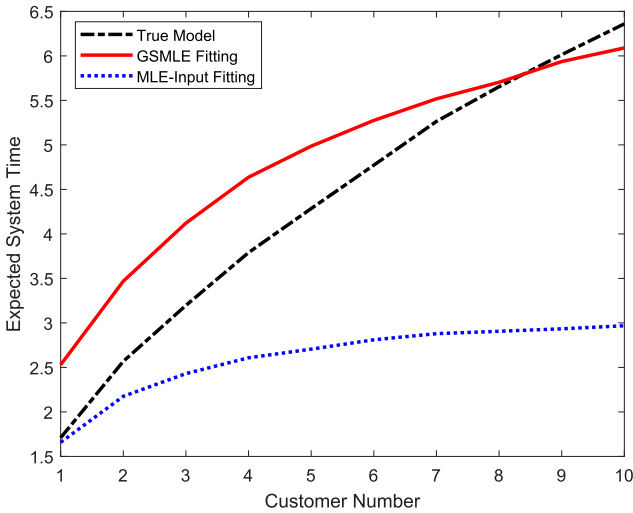
Figure 7. (Color online) The Expected System Times of the First 10 Customers in the True Model and M/M/1 Queueing Models Calibrated by GSMLE Fitting and MLE-Input Fitting, Respectively, Based on 10^4 Independent Experiments

Figure 8. (Color online) Trajectories of the GSMLE with 1,000 Observations Based on Sample Size $M = 10^4$ (Blue Line) and $M = 10^5$ (Red Line) and Known Observational Kernel (Black Line)
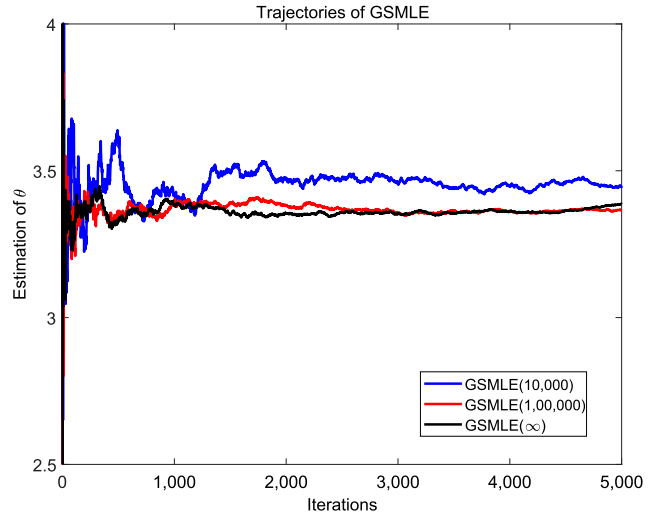

parameter $\theta = 3$, $\lambda_k = a/k$ with $a = 1$, starting point $\theta_0 = 2.5$, and feasible set $\Theta = [2,4]$. We show one trajectory of SA based on $T = 1,000$ observations, and $M = N = 10^4, 10^5$ simulated samples. For this example, we can also calculate the GSMLE based on the analytical forms of the transitional kernel and its derivatives and simulation from the true transition kernel for the hidden states, which is denoted as GSMLE(∞).

Figure 8 reports the trajectories of GSMLE for one set of 1,000 observations. We can see the trajectory of GSMLE(100,000) is very close to GSMLE(∞). With the initial point θ_0 uniformly distributed in Θ, GSMLE(∞) outputs 2.86 ± 0.05 (mean $\pm$ standard error) based on 100 macro experiments with 1,000 observations.

5. Conclusions

We provide a GSMLE using GLR estimators for the density and its derivatives, which allows a stochastic model without an analytical likelihood function to be directly fitted to the output data. Numerical experiments demonstrate that the GSMLE is flexible in

handling many types of stochastic models with different data-generating structures, including i.i.d. and Markovian/hidden Markov models and has potential to alleviate misleading results obtained by input fitting as illustrated by fitting a misspecified queueing model with output data, that is, the system times of the customers.

In future work, we plan to apply the GSMLE to calibration of complex stochastic models for decision-making problems with the underlying stochastic models driven by historic observations. We also plan to use it in likelihood ratio tests for hypothesis testing, the Akaike/Bayesian information criteria for model comparison, and Bayesian estimation for problems in which data are observed at the output level. Regarding optimization procedures, we investigate further techniques to handle the ratio bias in our log-likelihood derivative estimator, for example, letting the sample size sequentially go to infinity in the SA to get an asymptotically unbiased MLE and randomization techniques, such as in Rhee and Glynn (2015). Developing more efficient SA algorithms for the GSMLE utilizing higher-order derivatives of the likelihood and lower variance distribution sensitivity estimators are also interesting future directions. Finally, we also plan to study settings with online real-world data, with which applying GSMLE efficiently requires a joint sequential analysis with respect to both the real-world and simulated data as well as settings in which the Monte Carlo sampling is costly so that one would need to utilize their information most efficiently.

Table 1. Mean $\pm$ Standard Error of the Average System Time of the First 10 Customers Based on 10,000 Independent Experiments

	True model	GSMLE fitting	MLE-input fitting
Mean $\pm$ standard error	4.3 ± 0.04	4.8 ± 0.04	2.5 ± 0.02

References

Basawa IV, Bhat UN, Lund R (1996) Maximum likelihood estimation for single server queues from waiting time data. *Queueing Systems* 24(1–4):155–167.

Basawa IV, Bhat UN, Zhou J (2008) Parameter estimation using partial information with applications to queueing and related models. *Statist. Probab. Lett.* 78(12):1375–1383.

Bertsimas D, Gupta V, Paschalidis IC (2012) Inverse optimization: A new perspective on the Black-Litterman model. *Oper. Res.* 60(6):1389–1403.

Birge JR, Hortaçsu A, Pavlin JM (2017) Inverse optimization for the recovery of market structure from market outcomes: An application to the miso electricity market. *Oper. Res.* 65(4):837–855.

Cappé O, Moulines E, Rydén T (2005) *Inference in Hidden Markov Models* (Springer, New York).

Del Moral P (2004) *Feynman-Kac Formulae* (Springer, New York).

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B.* 39(1):1–38.

Doucet A (2001) *Sequential Monte Carlo Methods* (Wiley Online Library, Hoboken, NJ).

Dymarski P, ed., (2011) *Hidden Markov Models, Theory and Applications* (CRC Press, Boca Raton, FL).

Esfahani PM, Shafieezadeh-Abadeh S, Grani A Hanasusanto DK (2018) Data-driven inverse optimization with imperfect information. *Math. Programming* 167(1):191–234.

Fearnhead P (2004) Filtering recursions for calculating likelihoods for queues based on inter-departure time data. *Statist. Comput.* 14(3): 261–266.

Fu MC (2006) Gradient estimation. Henderson SG, Nelson BL, eds. *Handbooks in Operations Research and Management Science*, vol. 13 (Elsevier, Amsterdam), 575–616.

Fu MC (2015) Stochastic gradient estimation. Michael C, ed. *Handbooks of Simulation Optimization* (Springer, New York), 105–147.

Fu MC, Hong LJ, Hu J-Q (2009) Conditional Monte Carlo estimation of quantile sensitivities. *Management Sci.* 55(12):2019–2027.

Glasserman P (1991) *Gradient Estimation via Perturbation Analysis* (Kluwer Academic Publishers, Boston).

Glynn PW, Peng Y, Fu MC, Hu JQ (2020a) Computing sensitivity of distorted risk measure. *INFORMS J. Comput.* Forthcoming.

Glynn PW, Fan L, Fu MC, Hu J-Q, Peng Y (2020b) Technical note— Central limit theorems for estimated functions at estimated points. *Oper. Res.*, ePub ahead of print May 21, https://doi.org/10.1287/opre.2019.1922.

Goeva A, Lam H, Qian H, Zhang B (2019) Optimization-based calibration of simulation input models. *Oper. Res.* 67(5):1362–1382.

Heidergott B, Volk-Makarewicz W (2016) A measure-valued differentiation approach to sensitivity analysis of quantiles. *Math. Oper. Res.* 41(1):293–317.

Hong LJ (2009) Estimating quantile sensitivities. *Oper. Res.* 57(1):118–130.

Hong LJ, Liu G (2009) Simulating sensitivities of conditional value at risk. *Management Sci.* 55(2):281–293.

Hong LJ, Liu G (2010) Pathwise estimation of probability sensitivities through terminating or steady-state simulations. *Oper. Res.* 58(2): 357–370.

Jiang G, Fu MC (2015) Technical note—On estimating quantile sensitivities via infinitesimal perturbation analysis. *Oper. Res.* 63(2):435–441.

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. *J. Royal Statist. Soc. Ser. B. Statist. Methodology* 63(3):425–464.

Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications* (Springer, New York).

Larson RC (1990) The queue inference engine: Deducing queue statistics from transactional data. *Management Sci.* 36(5):586–601.

Lei L, Peng Y, Fu MC, Hu J-Q (2018) Applications of generalized likelihood ratio method to distribution sensitivities and steady-state simulation. *J. Discrete Event Dynamic Systems* 28(1): 109–125.

Liu G, Hong LJ (2009) Kernel estimation of quantile sensitivities. *Naval Res. Logist.* 56(6):511–525.

Liu JS, Chen R (1998) Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* 93(443):1032–1044.

Peng Y, Fu MC, Hu J-Q (2014) Gradient-based simulated maximum likelihood estimation for Lévy-driven Ornstein-Uhlenbeck stochastic volatility models. *Quant. Finance* 14(8):1399–1414.

Peng Y, Fu MC, Hu J-Q (2016) Gradient-based simulated maximum likelihood estimation for stochastic volatility models using characteristic functions. *Quant. Finance* 16(9):1393–1411.

Peng Y, Fu MC, Glynn PW, Hu J-Q (2017) On the asymptotic analysis of quantile sensitivity estimation by Monte Carlo simulation. Chan WKV, D'Ambrogio A, Zacharewicz G, Mustafee N, Wainer G, Page E, eds. *Proc. 2017 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 2336–2347.

Peng Y, Fu MC, Hu J-Q, Heidergott B (2018) A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Oper. Res.* 66(2):487–499.

Pickands J III, Stine RA (1997) Estimation for an M/G/∞ queue with incomplete information. *Biometrika* 84(2):295–308.

Rhee C-H, Glynn PW (2015) Unbiased estimation with square root convergence for sde models. *Oper. Res.* 63(5):1026–1043.

Ross, JV, Taimre T, Pollett PK (2007) Estimation for queues from queue length data. *Queueing Systems* 55(2):131–138.

Royden HL (1988) *Real Analysis*, 3rd ed. (Macmillan, New York).

Rudin W (1964) *Principles of Mathematical Analysis* (McGraw-Hill Education, New York).

Shao J (2003) *Mathematical Statistics* (Springer, New York).

Suri R, Zazanis MA (1988) Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Management Sci.* 34(1):39–64.

Tarantola A (2005) *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM, Philadelphia).

Van der Vaart AW (2000) *Asymptotic Statistics* (Cambridge University Press, Cambridge, UK).

Wang T-Y, Ke J-C, Wang K-H, Ho S-C (2006) Maximum likelihood estimates and confidence intervals of an M/M/R queue with heterogeneous servers. *Math. Methods Oper. Res.* 63(2):371–384.

Whitt W (1982) Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* 30(1):125–147.

Yijie Peng is an assistant professor in the Department of Management Science and Information Systems in the Guanghua School of Management at Peking University. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare.

Michael C. Fu holds the Smith Chair of Management Science in the Robert H. Smith School of Business with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland.

His research interests include simulation optimization and applied probability with applications in supply chain management and financial engineering. He is a fellow of INFORMS and IEEE.

Bernd Heidergott is a professor of stochastic optimization at the Department of Econometrics and Operations Research at the Vrije Universiteit Amsterdam, Netherlands. He is programme director of the BSc and MSc econometrics and operations research and research fellow of the Tinbergen Institute and of EURANDOM. His research interests are optimization and control of discrete event systems, perturbation analysis of Markov chains, max-plus algebra, and social networks.

Henry Lam is an associate professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research interests include Monte Carlo simulation and optimization under uncertainty.